This is the accepted version of the following article.

Sekine, K., & Kita, S. (2015/accepted). Development of multimodal discourse comprehension: Cohesive use of space by gestures. *Language, Cognition and Neuroscience*.

Running head: Integration of speech and gesture in discourse

Development of multimodal discourse comprehension: Cohesive use of space by gestures.

Kazuki Sekine[a,*] & Sotaro Kita[a]

[a] University of Warwick

* Corresponding author:

Kazuki Sekine

University of Warwick, Department of Psychology, Coventry CV4 7AL, United Kingdom Telephone; +442476523096

Email for Sekine; kazuki@tkc.att.ne.jp

[9641 words]

1st April, 2015

Abstract

This study examined how well 5-, 6-, 10-year-olds and adults integrated information from spoken discourse with cohesive use of space in gesture, in comprehension. In Experiment 1, participants were presented with a combination of spoken discourse and a sequence of cohesive gestures, which consistently located each of the two protagonists in two distinct locations in gesture space. Participants were asked to select an interpretation of the final sentence that best matched the preceding spoken and gestural contexts. Adults and 10-year-olds performed better than 5-year-olds, who were at chance level. In Experiment 2, another group of 5-year-olds were presented with the same stimuli as in Experiment 1, except that the actor showed hand-held pictures, instead of producing cohesive gestures. Unlike cohesive gestures, one set of pictures was self-explanatory, and did not require integration with the concurrent speech to derive the referent. With these pictures, 5-year-olds performed nearly perfectly and their performance in the identifiable pictures was significantly better than those in the unidentifiable pictures. These results suggest that young children failed to integrate spoken discourse and cohesive use of space in gestures, because they cannot derive a referent of cohesive gestures from the local speech context.

Key words: gesture, speech, discourse, multimodal communication, preschool children

Children typically learn their language in a multimodal environment. Given that gestures often convey information that is not conveyed in the accompanying speech (McNeill, 1992), and adults may sometimes produce utterances that are difficult for children to understand, gestures can potentially be an important source of information for children. To what extent children can benefit from gesture when speech is ambiguous is an important question. This study examined children's abilities to integrate information from gesture and speech at the discourse level to disambiguate discourse that is ambiguous when only speech is taken into account. Discourse is defined in this study as a structure in communication signals that spans over multiple sentences and multiple gestures.

Studies on gestures have revealed that during a narrative, an adult speaker creates coherent discourse not only by using linguistic devices but also by using idiosyncratic speech-accompanying gestures (Gullberg 2006; McNeill 1992, 2005; McNeill & Levy 1993; Yoshioka, 2005). Gestures can contribute to cohesion by repeating the same form of a hand shape or the same location where the gesture is produced to indicate continuity (McNeill, 1992). These gestures are called cohesive gestures, because they serve to "tie together thematically related but temporally separated parts of discourse"

(McNeill, 1992, p.16). In this study, we focus on gestural cohesion through the systematic use of locations in the space in front of the speaker, and examine how children and adults comprehend gestural cohesion and spoken discourse.

Gestures can assign a particular referent to a specific area in the space in front of the speaker. For example, when introducing a new protagonist in a narrative, adult speakers often assign them to a specific area by placing or directing a gesture to the area. When the same referent is mentioned again later, the same location is gesturally indicated (Gullberg 2006; So, Kita & Goldin-Meadow, 2009). In other words, once a location is assigned to a particular referent, it is often maintained throughout the discourse, not unlike the use of space for co-reference in sign language (e.g., Bellugi & Klima, 1982). Studies on the acquisition of sign language have shown that by age 4, deaf children comprehend that non-present referents are associated with locations in signing space, and that by age 6, they can assign the non-present referents in signing space themselves (Bellugi, Lillo-Martin, O'Grady, & van Hoek, 1990; Lillo-Martin,Bellugi, Struxness, & O'Grady, 1985).

Some deictic gestures serve purely cohesive functions, in that they indicate a location to which a referent should be assigned, without encoding information about the referent. For such gestures, the referent has to be inferred from the accompanying speech. For example, deictic gestures can assign a referent, specified in the speech, to a location in gesture space (Cassell, 1991; McNeill, 1992). Cohesive gestures are not necessarily pointing gestures with an extended index finger; sometimes the hand is used

as if to place an imaginary entity in gesture space (see Figure 1, Pictures 2-9). Sometimes iconic gestures (depicting action, motion and shape on the basis of similarity) are located in a specific area in gesture space (see Figure 1, Picture 11). This study investigated the comprehension of cohesive gestures that use space cohesively in discourse.

The referent of a pronoun in discourse can sometimes be ambiguous. However, previous studies have shown that adult listeners use grammatical function (Gordon, Grosz, & Gilliom, 1993) or gender of pronoun (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000) as a clue to disambiguate the referents of pronouns. Arnold et al (2000) also found that order-of-mention can be clues for listeners (Arnold et al., 2000). For example, English and Japanese speakers tend to interpret the referent of a pronoun (for English) or null pronoun/ zero marking (for Japanese), as being co-referential with the first mentioned character in a sentence describing more than one person (e.g., Goodrich & Hudson-Kam, 2012; Ueno & Kehler, 2010). However, when a full (or null) pronoun is used after describing more than one person with same gender, it is more difficult to identify the referent of the pronoun. But, in such cases, cohesive gestures accompanying the ambiguous pronoun help adult listeners to disambiguate the referent of the pronoun (Goodrich & Hudson-Kam, 2012, Furuyama, 2001).

Multiple component processes are necessary in order to interpret cohesive gestures. One process is *Local Cross-modal Referent Resolution*, in which the referent of gestures or words is clarified through information from the other modality. This process is required every time the listener encounters

a cohesive gesture. For example, this process may clarify the referent of a gesture, using the information from the concurrent words in speech. The important cue for such referent resolution is the synchronisation between speech and gesture. For example, when a pointing gesture is synchronized with the phrase "an old man" within the sentence “an old man and a boy went to a park”, then one would interpret the referent of the gesture to be an old man, not a boy. Note that information intrinsic to a single modality, for example, the form of a gesture, can also contribute to the referent resolution. For example, when a speaker is making his hands into a cupped shape while saying ‘a ball and a stick’, then one would interpret the referent of the gesture is likely to be a ball, not a stick. In this study, however, we focus on the cross-modal cue, that is, speech-gesture synchronization.

Another necessary process is *Spatial Mapping Management,* which has two subcomponents. First, the referent of a gesture is associated with a particular location in space that the gesture indicates. For example, when one has produced a gesture for “an old man” in a particular location, one would interpret the location to be associated with the old man. Second, this association between the gesture’s referent and the location is maintained over multiple utterances. For example, when the same referent re-appears in the narrative (“An old man and a boy went to a park. The old man was happy.”), and a pointing gesture that accompanies the second mention (“the old man”) indicates the same location as the pointing gesture with the first mention; then one would interpret the consistency of location as an indication of co-reference.

The output of the *Spatial Mapping Management* may further be used for *Local Cross-modal Referent Resolution* for speech. For example, when the referent is underspecified in the verbal utterance, (e.g. “He” in “An old man and a boy went to the park. He was happy.” could refer to either the old man or the boy). Cohesive use of space in gestures can disambiguate the referent of "he. A pointing gesture accompanying "he" may indicate the same location as the preceding pointing gesture that has accompanied "an old man". The spatial mapping between the location and the referent suggests that the referent of "he" is the old man.

In the current study, the key dependent variable was how often a listener-viewer successfully used cohesive gestures to disambiguate speech. When a listener-viewer failed to do so, then the listener-viewer must have failed in at least one of the three components: *Local Cross-modal Referent Resolution* for gesture, *Spatial Mapping Management* and *Local Cross-modal Referent Resolution* for speech. This study investigated whether the success or failure of the first component determined children’s performance.

Studies on the acquisition of sign language have shown that by age 4, deaf children comprehend that physically present referents are associated with locations in signing space, and that by age 6, they can assign the non-present referents in signing space themselves (Bellugi, Lillo-Martin, O’Grady, & van Hoek, 1990; Lillo-Martin,Bellugi, Struxness, & O’Grady, 1985). These findings indicate that deaf children can do *Spatial Mapping Management* by 6 years old. However, it is not clear whether hearing children can do *Local Cross-modal Referent Resolution* for gestures, because in sign language the

referent of signs can be derived from a single modality. Thus, in the current study, we focused on whether *Local Cross-Modal Referent Resolution*, particularly, for cohesive gesture (or cohesive use of a visual modality), is difficult for younger children.

Local Cross-Modal Referent Resolution and *Spatial Mapping Management* are necessary not only for cohesive gestures but also for other cohesive use of visual modality in discourse. For example, these processes are necessary when pictures accompany discourse in a spatially consistent way, to give extra information about multiple referents. Thus, these concepts are general and applicable to any visual cohesion in conjunction with verbal (spoken) discourse.

The study of comprehension of cohesive gestures in children is novel in the following way; most of the previous research on speech-gesture integration in comprehension focused on the processing of a single gesture at a time. Some studies have been shown that adults and children can pick up information conveyed solely in a gesture (e.g. Broaders & Goldin-Meadow, 2010; Goldin-Meadow & Sandhofer, 1999; Goodrich & Hudson-Kam, 2009; Kelly & Church, 1998; Namy, Cambell, & Tomasello 2004; Tomasello, Striano, & Rochat, 1999). Furthermore, gestures facilitate children's comprehension of semantically co-expressive words (McNeil, Alibali, & Evans, 2001; Morford & Goldin-Meadow, 1992). Other studies have shown how adults and children integrate gesture and speech so that each contributes unique information to the unified interpretation (adults: Cocks, Sautin, Kita, Morgan, & Zlotowitz, 2009; Kelly, Özyürek, & Maris, 2010; children: Kelly,

2001; Sekine, Sowden, & Kita, in press). Unlike the previous studies on children's comprehension of speech-gesture combinations, the current study investigated how children integrate the cohesive use of space in a sequence of gestures with spoken discourse.

Though no previous studies investigated children's comprehension of cohesive gestures, studies have shown that adult listener-viewers take up information from the cohesive use of space in gesture. McNeill and his colleagues (Cassell, McNeill, & McCullough, 1998; McNeill, Cassell, & McCullough, 1994) presented a video-recorded narrative to adult participants, who then re-told the story to a listener. In the stimulus narrative, the narrator set up two referents in the frontal space of the speaker with deictic gestures, and then linguistically referred back to one of the referents, but pointed to the wrong space (the space for the other referent) at the same time. When retelling the narrative, participants attempted to incorporate information from speech and gesture even when they were incongruent with each other. In a more recent study, Goodrich and Hudson-Kam (2012) examined whether pronoun interpretation is affected by cohesive gestures. In English, when a speaker introduces two same-gender protagonists with full nouns in a sentence and refers to one of them with a pronoun in the next sentence, a listener tends to interpret the ambiguous pronoun as the first-mentioned protagonist. This is called the first-mention bias. In their study, participants watched an actor narrating a short story about two protagonists with a sequence of cohesive gestures. It was found that when participants saw the narrator's cohesive gesture indicating the second-mentioned protagonist, while listening to an

ambiguous pronoun, they were more likely to go against the first-mention bias. That is, they tended to interpret the pronoun as the second-mentioned protagonist. Thus, these studies indicate that the adult participants derived information from the cohesive gestures. However, it is not clear whether children also have the ability to derive information from cohesive gestures. Furthermore, it is not clear what the component processes (such as, *Local Cross-modal Reference Resolution* for gestures, *Spatial Mapping Management*, *Local Cross-modal Reference Resolution* for speech) contribute to cohesive gestures' influence on discourse understanding.

We investigated whether Japanese-speaking 5-, 6-, 10-year-olds and adults can integrate spoken discourse and cohesive gestures and whether *Local Cross-modal Referent Resolution* for gesture is crucial in their success in the integration. The current study tested these age groups for the following reasons. First, children start using cohesive gestures that co-occur with spoken referential expressions and locate the referents in abstract space at around 8 years old (McNeill, 1992), and then use them frequently from 10 or 11 years old (Cassell, 1991; Sekine & Furuyama, 2010). This is why we included 10-year-olds in this study. Second, by age 5, children can integrate information from a short sentence and a single iconic gesture (Sekine et al., in press) or a single pointing gesture (Kelly, 2001) in a paradigm similar to the current study. The ability to integrate speech and a single gesture is a pre-requisite for integration of speech and a sequence of cohesive gestures. Third, until 4 years old, Japanese children tend to overuse zero-marking even when they introduced and re-introduced referents in a story (Clancy, 1992).

Thus, we decided to include children who are older than 5 years old in this study, because they start using zero-marking properly as a cohesive device from 5 years old. Fourth, children start formal education from 6-year-olds in Japan, in which they systematically learn about narratives in their school. Therefore, there may be a large difference between 5- and 6-year-olds.

In Experiment 1, we showed each of the participants in the four age groups video clips of an actor producing passages consisting of three sentences with accompanying cohesive gestures. Each passage referred to two protagonists. The first two sentences referred to the two protagonists by conjoined subject noun phrases, where two protagonists are connected by 'and' in a subject slot (e.g., "a boy and a girl" in "a boy and a girl are running."). The accompanying cohesive deictic gestures consistently assigned one protagonist to the right and the other to the left in the frontal space of the actor. The third sentence was ambiguous without an overt subject noun phrase and could refer to one of the protagonist's (or both protagonists') actions, but a cohesive gesture was produced in either the right or left space to depict the movement of a protagonist (e.g., a gesture in picture 11 in Figure 1 for "tumbling down"), and made it clear which protagonist (always only one) performed the action. Participants were asked to indicate which protagonist performed the action referred to in the third sentence in a forced choice task.

The third sentence with no overt subject noun phrase is grammatical in Japanese. It is common to omit an argument in Japanese especially when the information can be recovered from the context (Shibatani, 1990). The following is an example of Japanese discourse used in the current study.

1. Nori-kun-to Yuuto-kun-ga hodoukyou wo watasimasu

Nori-kun-and Yuuto-kun-NOM pedestrian bridge-ACC cross.PROG.Polite

"Nori-kun and Yuuto-kun are crossing a pedestrian bridge."

2. Nori-kun-to Yuuto-kun-wa kaidan wo nobotteimasu

Nori-kun-and Yuuto-kun-TOP stairs-ACC ascend.PROG.Polite

"Nori-kun and Yuuto-kun are ascending stairs."

3. Suruto totsuzen, korondeshimaimashita

and suddenly tumble.down-regrettably.Polite.PST

"and suddenly, tumbled down."

"Nori-kun" and "Yuuto-kun" are common Japanese boys' names. In each segment, the first line is the original Japanese speech, the second line shows the gloss (see Figure 1 for what each abbreviation stands for), and the third line is the English translation. Japanese does not have articles or commonly used third person pronouns (Shibatani, 1990). Thus, it is natural to use full noun phrases for maintained referents in the second sentence. The omitted subject in the third sentence, which is grammatical in Japanese, makes it ambiguous which protagonist(s) is mentioned. As mentioned above, the omitted subject can often be recovered from the context, such as accompanying cohesive gestures. Figure 1 shows how gestures were used in this study.

In Experiment 2, we tested whether the component process for integration, *Local Cross-modal Referent Resolution* for the visual modality

(e.g., gesture), is difficult for young children. Another group of 5-year-olds were presented with essentially the same stimuli as in Experiment 1, except that the actor showed hand-held pictures, instead of producing cohesive gestures. There were two types of hand-held pictures that differed in difficulty of *Local Cross-modal Reference Resolution*. The first type, Identifiable Pictures, represented protagonists that are unique and identifiable without information from concurrent speech (e.g, a picture of a frog and a picture of a mouse). For these pictures, *Local Cross-modal Reference Resolution* is not necessary. The second type, Unidentifiable Pictures, represented protagonists that are not identifiable without information (proper names) from speech (e.g., a picture of a boy in long-sleeve shirt and a picture of a boy in a short-sleeve shirt, who were referred to with proper names, "Nori-kun" and “Yuuto-kun”, respectively in speech) (See Figure 1). For these pictures, *Local Cross-modal Reference Resolution* for the visual modality is necessary.

Experiment 1

Method

Participants

24 5-year-olds (mean age: 5;03, range: 4;10 to 5;09), 24 6-year-olds (mean age: 6;03, range: 5;12 to 6;09), 24 10-year-olds (mean age: 10;03, range: 9;10 to 10;08), and 24 adults (mean age: 23, range: 18 to 31), who were all monolingual speakers of Japanese participated. Each age group had 12 females and 12 males.

Material

An actor was filmed producing a combination of gestures and a short passage[2]. In total, 17 vignettes were made (two for practice and 15 for the main experiment). The lower part of the actor's face was covered by a mask to conceal lip movements (see Figure 1). The speech spoken through the mask was not available to participants at all. The original speech was dubbed by the recorded speech for stimuli. Each vignette consisted of three short sentences and gestures. The three sentences will firstly be described. In the first sentence, the actor introduced two protagonists by two conjoined full nouns or proper names in the subject positions and described an event involving them. In the second sentence, she referred to the same two protagonists by two conjoined full nouns or proper names in the subject positions again. She also referred to an event in the sentence. In the third sentence, she described one protagonist's action as a result of the event, but omitted the subject. Thus, participants could not know which character performed the action if they took only the speech into account. Due to the characteristic Japanese discourse, the third sentence has only a verb (no noun phrases). Thus, it is more natural if the gesture depicts the action referred to by the verb (pointing may have been more natural if there was an overt noun phrase).

As mentioned, the three sentences were accompanied by cohesive gestures performed by the actor. In the first sentence, gestures were produced to assign each of the two protagonists to the actor's right and left sides of frontal space with her right and left hand respectively, when each protagonist was mentioned, as if she places two entities in the space ((2)-(5) in Figure 1).

In the second sentence, two further gestures placed the protagonists in the same locations as in the first sentence ((6)-(9) in Figure 1). The actor's hands were held in the air at the beginning of the third sentence ((10) in Figure 1). In the third sentence, either her right or left hand depicted one of the protagonists' actions within the right or left space, respectively. The stationary hand was held until the other hand finished the gesture ((11) in Figure 1). In other words, the gesture specified which protagonist performed the action. Finally, both hands were replaced in the actor's lap.

The gesture that had represented one protagonist was held in the air while introducing or mentioning the other protagonist. Our assumption was that post-stroke hold would help participants to maintain the association between the location and the referent and clarify the contrast between the locations for each protagonist. If the actress had put her hand back to her lap after each gesture stroke (without a hold), such scaffolding is not available. Thus, post-stroke hold would provide young children the best chance to succeed in the task. This use of a gestural hold was attested in gestures spontaneously produced by adults during narrative (Sekine & Kita, under review).

Of the fifteen main vignettes, seven of them had the actor placing the first-mentioned protagonist in the first two sentences on the right (and the second-mentioned protagonist on the left) and the remaining eight had the actor placing the first-mentioned protagonist on the left (and the second on the right). For each vignette, we made four versions to counterbalance the location of the gestures: the location (left or right) in which a gesture

assigned the first protagonist in the story, and the location (left or right) in which a protagonist's action was depicted in the third sentence. The order that the two protagonists were introduced in the actor's script was fixed in each story. Thus the speech was identical across the four versions for each video. Each video lasted about 20 seconds. See Figure 1 for an example.

"Insert Figure 1 about here"

The third sentence in each vignette did not have an overt subject. It is grammatical in Japanese to omit arguments of a sentence (Shibatani, 1990). As Japanese does not have subject-verb agreement (e.g., based on number and person), it is not clear from the speech whether protagonist A or protagonist B or both protagonists performed the action. However, the accompanying gesture disambiguated who performed the action. Thus, participants needed to gain information from the accompanying gestures to get the correct answer. As mentioned above, the use of the full noun phrases in the second sentence is natural in Japanese. This is because Japanese does not use third person

pronouns in everyday discourse (third person pronouns are used mainly in translations from European languages) and omitting an overt subject in the second sentence would have made the story too unclear. Thus, as Clancy (1992) described, the major referential choice in Japanese discourse is between ellipses versus nominal reference. Because Japanese has no subject-verb agreement, no information about the omitted subjects is recoverable from the verb[3]. Japanese speakers must rely heavily on the listener's ability to interpret the referent of omitted arguments from the context. Although few experimental studies on interpretation of Japanese discourse have been conducted, one such study found that an ambiguous referent in Japanese narrative is sometimes disambiguated by cohesive gestures (Furuyama, 2001).

After the video stimulus was presented, participants were asked to pick an answer from three choices. In the analysis, the three choices were labeled as follows: *correct choice*, *incorrect-protagonist choice*, and *both-protagonists choice*. In case of the example in Figure 1, after participants watched the clip, the experimenter asked the participants "Did Nori-kun tumble down, did Yuuto-kun tumble down, or did both of them tumble down?" The cohesive gesture in the third sentence was produced to depict the movement of a protagonist in the space associated with Yuuto-kun in the first two sentences. Therefore, the answer that Yuuto-kun tumbled down was coded as a *correct choice*. The answer that Nori-kun tumbled down was coded as an *incorrect-protagonist choice*. The answer that both tumbled down was coded as *both-protagonists choice* (also an incorrect choice).

Procedure

Participants were tested individually. Participants were asked to watch a video stimulus embedded in a PowerPoint presentation on a laptop with a 15 inch screen. Before watching each vignette, an experimenter told the child participants what protagonists would appear in the next vignette to make it easier for children to remember the protagonists. After each vignette, participants were asked to pick one of three choices about who performed the action in the third sentence. Regardless of which option they picked, the experimenter gave them positive feedback after each trial, such as "well done" or "good job". Two practice trials were followed by 15 experimental trials. Each participant was presented with one of the eight counterbalanced sets for the experimental trials. Participants were presented with the stories in one of the two fixed orders (one order was the reverse of the other). Thus, there are a total of eight counterbalancing sets: four gesture locations (as described in the materials section) in either of two vignette orders (forward vs. backward). The experiment lasted approximately 10 minutes.

Results

The proportion of correct choices did not significantly differ between the eight counterbalancing sets. In addition, the proportion of correct choices did not differ between when the first-mentioned protagonist was the target protagonist and when the second-mentioned protagonist was the target protagonist for each age group. Thus, the data collapsed across counterbalancing sets and the order in which the protagonists were introduced

in a story.

Correct choices for each age group

We examined whether information from gesture influenced the participant's choice of a target protagonist. First, we examined whether proportions of trials in which participants selected the target (correct) protagonist that was indicated by the location of the gesture in the third sentence were higher than chance level (.50). We conducted this analysis for both sides; when the target protagonist was located by a gesture on the right side and when it was located on the left site. If participants' choices were not influenced by gesture at all, the proportion of trials with a correct choice should be at chance. We excluded trials with the both-protagonists choice from this analysis because participants did not select this choice very often (0 - 13% of the trials, depending on the age group; see the second row of Table 2).

"Insert Table 1 about here"

A Wilcoxon Signed-ranks test indicated that the proportions of trials with the correct choice were significantly higher than chance level (0.5) for

all age groups except 5-year-olds for both sides (Table 1). Thus, for both right side and left side, the information from cohesive gestures influenced the choice of the referent for the elided subject in the third sentence for all age groups except 5-year olds.

Next, we examined age differences in overall accuracy, that is, the ability to use cohesive gestures to disambiguate the third sentence in the stimulus discourse. As noted above, the information encoded in cohesive gestures was necessary to select the correct choice. A Kruskal-Wallis test was conducted to evaluate differences among four age groups on the mean proportion of trials with a correct choice (see Figure 2 for the means and SEs). The test was significant, $\chi^2(3, N = 96) = 47.87$, $p < .001$, *Cramer's* V = .71. Post hoc comparisons (Mann-Whitney tests with Bonferroni correction) showed that adults chose the correct answer significantly more often than 5- and 6-year-olds did, and that 10-year-olds selected the correct answer significantly more often than 5-year-olds did. This indicates that it is relatively difficult for 5- and 6-year-olds to integrate information from both cohesive gestures and spoken discourse.

"Insert Figure 2 about here"

Proportion of correct choices between the two one-protagonist choices

We calculated the mean proportion of trials with each type of error for each age group (Table 2). It turned out that the participants rarely selected the *both-protagonists choice*. This is perhaps not surprising as the key gesture in the third sentence was produced by just one hand. In addition, most adults did not make any errors.

"Insert Table 2 about here"

As the participants rarely picked the *both-protagonists choice*, we examined whether the proportion of correct choices was above the chance level (50%), when they picked one of the two one-protagonist choices (*correct choice* vs. *incorrect-protagonist choice*) (Table 3), by excluding trials in which the participants picked the both-protagonist choice. Some participants were excluded from this analysis because they selected the both-protagonist choice in all trials. Note that the chance level was 50% in this analysis because of the counterbalancing of gesture locations; for a given story, the correct referent of the third sentence was the first-mentioned protagonist for half of the participants, and the second-mentioned protagonist for the other half of the participants.

A Wilcoxon Signed-ranks test indicated that the proportions of trials

with the correct choice were significantly higher than chance level (0.5) for all age groups except 5-year-olds (Table 3). This indicated that it was difficult for 5-year-olds to integrate information from spoken discourse and cohesive gestures and pick the correct protagonist.

"Insert Table 3 about here"

Possible response biases in 5-year-olds.

Lastly we examined response biases in 5-year-olds. There is no evidence that they had a bias to choose the first mentioned protagonist or second mentioned protagonist. The result showed that the mean proportion of responses selecting the first mentioned protagonist (M = 0.46, SD = 0.18) did not significantly differ from that of responses selecting the second mentioned protagonist (M = 0.48, SD = 0.18). (The proportions did not add up to 1 because they sometimes picked a both protagonist choice).

Similarly, there is no evidence that they had a bias to choose the protagonist established on the right side or left side (regardless of the space indicated by the gesture in the third sentence). The result showed that the mean proportion of responses selecting the right side protagonist (M = 0.45, SD = 0.18) did not significantly differ from that of responses selecting the left side protagonist (M = 0.49, SD = 0.19).

Discussion

Experiment 1 tested how well children and adults integrated information from cohesive gestures and spoken discourse. There are two main findings. First, we found no evidence that 5-year-olds could integrate information from spoken discourse and the cohesive gestures, but 6-year-olds could perform above chance level, although not as well as adults. Previous studies (Kelly, 2001; Sekine et al., in press) showed that when participants were shown video recordings of combinations of a single sentence and a single gesture, 5-year-olds could select correct choices above chance level. Thus, discourse-level integration of speech and gesture develops later than the single-sentence/gesture level integration.

Second, we provide evidence that adults integrate information from spoken discourse and cohesive gestures. This finding is in line with previous studies conducted in English, which showed that adult listeners take into account information conveyed by a speaker's gestures that are anaphorically used (Cassell et al., 1998; Goodrich & Hudson-Kam, 2012; McNeill et al., 1994).

Unlike adults and older children, 5-year-olds showed difficulty in integrating information from cohesive gestures and spoken discourse. As discussed in the introduction, the poor performance of 5-year-olds indicates their failure in one of the three component processes: *Local Cross-modal Reference Resolution* for gestures, *Spatial Mapping Management*, *Local Cross-modal Reference Resolution* for speech. We hypothesized that the first

component may be the key difficulty for 5-year-olds for the following reasons: the literature on deaf children's use of cohesive use of space in their signing indicates that 4-year-olds can do *Spatial Mapping Management* (e.g., Lillo-Martin, 1999; Lillo-Martin et al., 1985), and the literature on pronoun resolution suggests that children 5-year-olds may be able to do *Local Cross-modal Referent Resolution* for speech. This is because children as young as 3 years old can use various contextual cues (e.g., genders of protagonists or saliency of referent) to identify the referent of an ambiguous pronoun (Arnold, Brown-Schmidt, & Trueswell, 2007; Pyykkönenab, Matthewsc & Järvikivid, 2010). However, no previous studies investigated whether children can identify the referent of pointing (deictic) gestures, using information from concurrent speech. Thus, in Experiment 2, we investigate whether *Local Cross-modal Reference Resolution* for visual modality (such as gesture and picture that accompany discourse) was the stumbling block for the 5-year-olds in Experiment 1.

Experiment 2

Experiment 2 tested whether *Local Cross-modal Referent Resolution* for visual modality (including gestures) was posing difficulties for 5-year-olds in Experiment 1, by manipulating the difficulty of this processing component by replacing cohesive gestures with hand-held pictures and using two different types of pictures.

The stimuli were the same as Experiment 1 except that the gestures were replaced by hand-held pictures. The hand-held pictures consisted of two types

(Identifiable and Unidentifiable Pictures), and they differed in the difficulty of *Local Cross-modal Reference Resolution*. For the Identifiable Pictures, which represented a protagonist that is unique and identifiable without information from concurrent speech, *Local Cross-modal Reference Resolution* is not necessary. For the Unidentifiable Pictures, which represented a protagonist that is not indefinable without protagonist's proper names in speech, *Local Cross-modal Reference Resolution* is necessary. However, the reference resolution may be easier than cohesive gestures because the pictures provide a concrete image of the referents.

If 5-year-olds perform better in the items with the Identifiable Pictures than in the items with the Unidentifiable Pictures, that would indicate a problem in *Local Cross-modal Referent Resolution* for the visual modality. This would, in turn, suggest that 5-year-olds in Experiment 1 had a difficulty in deriving the referent of the gesture from concurrent words in speech

Method

Participants

The participants were 24 5-year-olds (mean age: 5;04, range: 5;00 to 5;09; 12 females), who were all monolingual speakers of Japanese and did not participate in Experiment 1.

Material and Procedure

We created hand-held pictures depicting each protagonist. The pairs of protagonists (e.g., a dog and a cat) for each story were the same as Experiment 1. Out of the two stories used in practice trials in Experiment 2, one had a pair

of Identifiable Pictures, and the other story had a pair of Unidentifiable Pictures. Out of the fifteen stories in the main trials, seven stories consisted of a pair of Identifiable Pictures (e.g., a mouse and a frog) and eight stories consisted of a pair of Unidentifiable Pictures (e.g., a boy in a long-sleeve shirt and a boy in a short-sleeve shirt, see Figure 3). As described above, the protagonists in the Unidentifiable Pictures were not indefinable without information (proper names) from speech (e.g., Nori-kun and Yuuto-kun; common Japanese boys' names), whereas the protagonists in the Identifiable Pictures could be identified without speech. Each protagonist was depicted on one side of a piece of cardboard (15cm × 21cm) in color, and the back side was blank. A stick was attached to the cardboard so that the actor could hold it and flip from the picture side to the blank side. The lower part of the actor's face was covered by a mask to conceal lip movements as in Experiment 1.

The vignettes were the same as Experiment 1 except that the gestures were replaced by hand-held pictures. We moved the pictures in such a way that relevant information was presented in an analogous way to gesturing in Experiment 1. We will illustrate this with the timing of movement of the actress' right hand (on the left in Figures 1 and 3). The hand held picture showed the picture-side and the gesture (Experiment 1) showed a downward stroke while the speaker uttered the noun phrase referring to a relevant protagonist in the first sentence (e.g., "Nori-kun" (the name) in line 1 of the text, uttered between Panels 2 and 3, in Figure 1 and Figure 3). This timing encouraged participants to identify the referent of the gesture and the picture with the referent of the noun phrase. Subsequently, the hand-held picture was

flipped and held with the blank-side facing forward and the gesturing hand (Experiment 1) was held in the air (where the gesture stroke ended). This hold phase encoded only the information as to where the protagonist was placed in preceding expression. The hold continued until the same protagonist was mentioned again in the second sentence (between Panels 6 and 7 in Figure 1 and Figure 3), during which the picture-side was shown and the gesture (Experiment 1) showed a downward stroke again. Then, the hand-held picture with the blank side facing forward and gesture (Experiment 1) were held until the end of the third sentence. In the other half of the stimuli, the actress' right hand moved in an arc downward and away from the midline during the third sentence to depict the movement of the protagonist, similarly to the left-hand movement between Panels 10 and 11 in Figure 1 and Figure 3). The hand-held picture was showing the blank side during this movement in the third sentence (see Panel 11, Figure 3).

"Insert Figure 3 about here"

As in Experiment 1, participants could not know which protagonist moved in the third sentence if they took only the speech into account. The procedure and the counterbalancing of items across participants were also identical to that in Experiment 1.

Result

Just like Experiment 1, the pattern of responses did not statistically

differ between the eight counterbalancing sets. Thus, the data were collapsed across counterbalancing sets.

First, we compared the proportion of trials with a correct choice between the two types of items. A Wilcoxon Signed-ranks test indicated that the proportion of correct choice in the Identifiable items (M = 0.99, SD = 0.03) was significantly higher than that in the Unidentifiable items (M = 0.86, SD = 0.14), $Z = 3.30$, $p < .001$, $r = 0.67$.

Next, we examined whether the proportion of trials with correct choices was above the chance level in the Identifiable and the Unidentifiable items. As none of the 5-year-olds in Experiment 2 selected the both-protagonist choice in any of the trials, we set a stringent chance level at .5, just as in Experiment 1. A Wilcoxon Signed-ranks test indicated that the proportions of the correct choice were significantly higher than chance level for both the Unidentifiable items, $Z = 4.25$, $p < .001$, $r = 0.87$, and the Identifiable items, $Z = 4.81$, $p < .001$, $r = 0.67$.

Discussion

The 5-year-olds showed better performance in the Identifiable Pictures than the Unidentifiable Pictures. The Unidentifiable Pictures put a higher cognitive load on the *Local Cross-modal Reference Resolution* for the visual modality. Thus, this suggests that the difficulty in *Local Cross-modal Reference Resolution* for the visual modality may be one of the reasons why the 5-year-olds in Experiment 1 failed to integrate speech and cohesive gestures. In addition, unlike Experiment 1, we found that the 5-year-olds

performed well above chance level in both Identifiable and Unidentifiable items. From this result, we can rule out another less interesting explanation for the result in Experiment 1, that 5-year-olds simply did not understand the task or instruction. This is because the task in Experiment 2 had comparable task structure and procedure as the task in Experiment 1.

The direct comparison between Experiment 1 and Experiment 2 may reveal the effect of different types of information in the visual modality. However, this deviates from the main purpose of the study, thus the results are in Supplementary Material.

General discussion

This study examined how well Japanese-speaking children and adults integrated information from spoken discourse and cohesive gesture in comprehension, and whether *Local Cross-Modal Referent Resolution* for gestures is difficult for younger children. There are two main findings. First, adults can successfully integrate spoken and gestural contexts to derive the correct interpretation, but this was difficult for 5-year-olds to do so. The performance improved with age, and 6- and 10-year-olds performed above chance-level, though not as well as adults. This indicates that children aged 6 years and older could derive the referents of cohesive gestures from speech, and use the meaning assigned to the locations by the gestures to disambiguate a semantically underspecified sentence.

Second, 5-year-olds have a difficulty in *Local Cross-modal Referent Resolution* for gestures. 5-year-olds performed at chance level, and worse

than other age groups with cohesive gestures in Experiment 1, where *Local Cross-modal Referent Resolution* for iconic gestures was necessary. Crucially, in Experiment 2, they performed worse for Unidentifiable items (requiring *Local Cross-modal Referent Resolution* for the visual modality) than for Identifiable items. These results indicate difficulty in *Local Cross-modal Referent Resolution* for the visual modality, including cohesive gestures, contributed to their difficulty in comprehension of cohesive gestures.

We attributed children's poor performance with cohesive gestures in Experiment 1 to their difficulty with *Local Cross-modal Referent Resolution* for gestures. This conclusion assumes that Experiments 1 and 2 have comparable levels of difficulty in *Spatial Mapping Management*. This assumption is based on the fact that cohesive gestures and picture presentation in the first and second sentences both involved small movements (a downward stroke for gestures and flipping for pictures), and very similar iconic movements in the third sentence (e.g., the hand or the handheld picture, showing the blank side drew an arc going laterally and downward to indicate "falling down"). However, it is important to acknowledge that the movements, especially those accompanying the first and second sentences, were not identical. Thus, this difference may have made *Spatial Mapping Management* in one experiment easier than the other. This concern is somewhat mitigated by the contrasting findings between the current study and the study of cohesive (anaphoric) use of space in sign language, as discussed below. Differences in relevant movements may be less of a concern for signs and gestures than for gestures and pictures in this Experiment. The comparison of

the results in the current study and in the sign language literature further supports the idea that 5-year-olds are not skilled at integrating cross-modal information. Studies on the acquisition of sign language have shown that by the age of 5, deaf children comprehend that signs can associate non-present referents with the locations in signing space (Lillo-Martin, 1999; Lillo-Martin et al., 1985). This finding tells us that if only one modality is used, 5-year-olds can do *Spatial Mapping Management* and resolve a referent. In contrast, with co-speech gestures of the type used in the current study, hearing children have to use information from speech to derive the referent of cohesive gestures, and they did not perform well in the task. Thus, we infer that 5-year-olds found *Local Cross-modal Referent Resolution* for gestures difficult because it required cross-modal integration of information.

This study focused only on *Local Cross-modal Referent Resolution* for gestures, and we cannot rule out the possibility that the other two components (*Spatial Mapping Management* and *Local Cross-modal Referent Resolution* for speech) also contribute to age-related differences in performance. Thus, future studies need to be designed to isolate the impact of the three components by manipulating the one of the three components, whilst the other two components remain identical.

The interpretation based on *Local Cross-modal Referent Resolution* for gestures explains why 5-year-olds succeeded in integrating information from speech and other types of gestures whose referents are easier to resolve. For example, Kelly (2001) showed that 5-year-olds correctly interpreted a spoken indirect request when an accompanying deictic gesture indicated a concrete

and visible object related to the request. The referents of such pointing gestures are easier to identify than cohesive gestures indicating abstract locations in physically empty space, as in Experiment 1, because concrete pointing gestures indicate the referent by means of spatiotemporal contiguity to the object. Sekine et al. (in press) found that 5-year-olds successfully integrated an aspect of an action described in a sentence and a different aspect of the action depicted in an accompanying iconic gesture to form a unified interpretation. The referents of these iconic gestures are easier to identify than the cohesive gestures that indicate locations, as in Experiment 1. This is because iconic gestures represent the referent (e.g., action, motion and shape) on the basis of similarity between the form/movement of the gesture and the part of referent, that is, gestural movement itself encodes certain properties of the referent. In contrast, cohesive gestures themselves do not encode properties of the referent, and the listener-viewer has to integrate information from the concurrent speech and from the memory of what referent had been associated with the indicated location. Thus, the similarity in iconic gestures reduces the need for cross-modal integration to resolve the referents, not unlike the Identifiable pictures in Experiment 2.

Significant difference of the proportions of trials with the correct choice was found between 5- and 6-year-olds in Experiment 1. This may be related to general development in the ability to link elements in discourse. Peterson and McCabe (1983) have analyzed young children's personal event narratives, and found that by age 5, children can tell a sequence of events in oral narratives, but they tend to dwell on a climactic event. However,

6-year-olds can make a well-formed story that orients a listener to who, what, and where something happened with some sort of climax. In other words, 6-year-olds were able to link various key elements of narrative in a coherent way. Such abilities in 6-year-olds might have allowed them to link discourse information conveyed by gesture and speech in the stimuli, and led them to perform better than 5-year-olds.

When we consider the findings from the current and previous studies on discourse and gesture production together, it seems that comprehension develops considerably earlier than production for cohesive gestures. Previous studies revealed that by approximately 9 or 10 years old, most children can flexibly use spoken referential expression to anaphorically identify referents in their narratives (Jisa, 2000; Karmiloff-Smith, 1985), and often gesturally locate the referents in abstract space, but the frequency does not yet attain adult level (Cassell, 1991; Sekine & Furuyama, 2010). Some studies also showed that children younger than 9 years old rarely produced cohesive gestures (McNeill, 1992; Sekine & Furuyama, 2010). In contrast, the current study showed that 6-year-olds can comprehend cohesive gestures above chance and 10-year-olds can do so at the adult level. Thus, we concluded that the comprehension of cohesive gestures develops earlier than the production of them.

There are two questions for future studies. The first important open question is whether our finding from the current study can be generalized to other languages. Given zero anaphora marking is used in Japanese discourse instead of personal pronouns, it would be interesting to investigate languages

that predominantly use personal pronouns, such as English. When designing experiments with other languages, stimulus vignettes should be constructed with language-specific considerations. For example, in English, one may use a pronoun with ambiguous referents (e.g., "he" with two male protagonists). The second question is about the effect of post-stroke hold in our stimuli on gestural contribution to discourse comprehension. We included post-stroke holds based on the assumption that the held hand would help maintain the association between the location and the referent (*Spatial Mapping Management*), and contrast the two locations in gesture space with different meanings (*Local Cross-modal Referent Resolution*). The future study should test this assumption. An alternative possibility is that the post-stroke hold may hinder the listener's comprehension as the hold phase alongside the other stroke phase provide the listener with too much information to process.

In summary, this study revealed that 5-year-olds have difficulty in using discourse cohesion established by cohesive gestures. We have argued that this is because they have difficulty in deriving the referents of cohesive gestures from concurrent words in speech. Children have to learn how to assign meaning to a location in abstract space indicated by cohesive gesture, using the meaning of concurrent speech. This ability develops during the elementary school period.

References

Arnold, J. E., Brown-Schmidt, S., & Trueswell, J. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and Cognitive Processes*, 22, 527-565.

Arnold, J. E., J. Eisenband, S. Brown-Schmidt & J. Trueswell. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eye tracking. *Cognition*, 76. B13-B26.

Bellugi, U., & Klima, E. (1982). From gesture to sign: Deixis in visual-gestural language. In R. J. Jarvella and W. Klein (Eds.), *Speech, place and action: Studies in deixis* (pp.297-313). New York: John Wiley & Sons.

Bellugi, U., Lillo-Martin, L., O'Grady, K., & van Hoek, K. (1990). The development of spatialized syntactic mechanisms in American Sigh Language. In W. H. Edmondson & F Karlson (Eds.) *The Fourth international symposium on Sign Language Research* (pp. 16-25). Hamburg: Signum-Verlag Press.

Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science, 21*(5), 623-628.

Boyatzis, C. J. & Watson, M. W. (1993). Preschool children's symbolic representation of objects through gestures. *Child Development*, *64*, 729-735.

Cassell, J. (1991). *The development of time and event in narrative: Evidence from speech and gesture*. Unpublished doctoral dissertation, University of Chicago.

Cassell, J., McNeill, D., & McCullough, K.E. (1998). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics and Cognitioni, 7*(1): 1-33.

Clancy, P. (1992). Referential strategies in the narratives of Japanese children, *Discourse Processes*, 15(4), 441-467.

Cocks, N., Sautin, L., Kita, S., Morgan, G., & Zlotowitz, S. (2009). Gesture and speech integration: An exploratory study of a man with aphasia. *International Journal of Language and Communication Disorders*, *44*, 795-804.

Furuyama, N. (2001), *De-syntactizing the theories of reference maintenance from the viewpoint poetic function of language and gesture: A case of Japanese discourse*, Unpublished doctoral dissertation, University of Chicago.

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science, 2*(1), 67-74.

Goodrich, W., & Hudson-Kam, C. L. (2009). Co-speech gesture as input in verb learning. *Developmental Science, 12*(1), 81-87.

Goodrich, W., & Hudson-Kam, C. K. (2012). Knowing 'who she is' based on 'where she is': The effect of co-speech gesture on pronoun comprehension. *Language and Cognition*, 4-2, 75-98.

Gordon, P. C., B. J. Grosz & L. A. Gilliom. (1993). Pronouns, names and the centering of attention in discourse. *Cognitive Science,* 17, 311–347.

Gullberg, M. (2006). Handling discourse: Gestures, reference, tracking, and communication Strategies in early L2. *Language Learning*, *56*(1), 155 -196.

Jisa, H. (2000). Increasing cohesion in narratives: A developmental study of

maintaining and reintroducing subjects in French. *Linguistics*, *38*(3), 591-620.

Karmiloff-Smith, A. (1985). Language and cognitive processes from a developmental perspective. *Language and Cognitive Processes*, *1*(1), 61-85.

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*, 325-349.

Kelly, S. D. & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development*, 69, 85-93.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21, 260-267.

Lillo-Martin, D. (1999). Modality effects and modularity in language acquisition: The acquisition of American Sign Language.

Lillo-Martin, D., Bellugi, U., Struxness, L., & O'Grady, M. (1985). The acquisition of spatiality organized syntax. *Paper and Report on Child Language Development*, 24, 70-78.

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2001). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2), 131-150.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago

Press.

McNeill, D., Cassell, J., & McCullough, K. E. (1994). Communicative effects of speech-mismatched gestures. *Language and Social Interaction*, *27*, 223-237.

McNeill, D., & Levy, E. T. (1993). Cohesion and gesture. *Discourse Processes*, *16*(4), 363-386.

Morford, M., & Goldin-Meadow, S. (1992). Comprehension and production of gesture in combination with speech in one-word speakers. *Journal of Child Language*, *19*(3), 559-580.

Namy, L. L., Campbell, A. L., & Tomasello, M. (2004). The changing role of iconicity in non-verbal symbol learning: A U-shaped trajectory in the acquisition of arbitrary gestures. *Journal of Cognition and Development, 5*(1), 37-57.

Pyykkönen, P., Matthews, D., & Järvikivi, J. (2010). Three-year-olds are sensitive to semantic prominence during online language comprehension: A visual world study of pronoun resolution. *Language and Cognitive Processes*, 25, 115-129.

Peterson, C., & McCabe, A. (1983). *Developmental psycholinguistics*: *Three ways of looking at a child's narrative*. New York: Plenum

Riggs, J. K., McTaggart, J., Simpson, A., & Freeman, R. P. J. (2006). Changes in the capacity of visual working memory in 5- to 10-year-olds. *Journal of Experimental Child Psychology*, 95, 18-26.

Tomasello, M., Striano, T., & Rochat, P. (1999). Do young children use objects as symbols? *British Journal of Developmental Psychology*, *17*,

563-584.

Sekine K., & Furuyama, N. (2010). Developmental change of discourse cohesion in speech and gestures among Japanese elementary school children. Rivista di psicolinguistica applicata, 10(3), 97-116.

Sekine, K., & Kita, S. (under review). The parallel development of the form and meaning of two-handed gestures and linguistic information packaging within a clause in narrative.

Sekine, K., Sowden, H., & Kita, S. 5-year-olds, but not 3-year-olds, semantically integrate information in speech and iconic gesture in comprehension. in press.

Shibatani, M. (1990). The languages of Japan. Cambridge: Cambridge University Press.

So, W. C., Kita, S., & Goldin-Meadow, S. (1990). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33, 115-125.

Ueno, M., & Kehler, A. (2010). The interpretation of null and overt pronouns in Japanese: Grammatical and pragmatic factors. In S. Ohlsson & R. Catrambone (eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. *Cognitive Science Society*, 2057-2062.

Wilkin, K., & Holler, J. (2011). Speakers' use of 'action' and 'entity' gestures with definite and indefinite references. In G. Stam, & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 293-308). Amsterdam: John Benjamins.

Yoshioka, K. (2005). *Linguistic and gestural introduction and tracking of*

referents in L1 and L2 discourse. Nijmegen, Radboud University.

Footnote

[1] Similar type of gestures has been suggested by gesture researchers, such as *abstract deixis/ pointings*" (McNeill, 1992, 2005), *referential gestures* (Gullberg, 2006; Yoshioka, 2005), and *entity gestures* (Wilkin & Holler, 2011). The definitions of their gestures include the property of gesture that do not represent any existing space, and that rather creates spaces and refer to locations that do not physically exist. However, they are different each other in terms of the repetitive quality, the usage in a discourse, the shape of gesture, type of speech accompanied with gesture.

[2] We selected passages by conducting a pre-test. In the pretest, we presented a written questionnaire, consisting of 20 candidate passages and forced-choice questions, to 20 adult native speakers of Japanese who did not participate in the main experiment, and asked them to pick a correct answer from same three choices as the main experiments. Based on the result, we excluded passages in which more than 60% of the adults picked a particular choice. We finally selected 15 passages for the main experiment. The 15 messages are passages where adults picked the *protagonist A choice (first-mentioned protagonist)*, *protagonist B choice (second- mentioned protagonist)*, and *both protagonists choice* in the forced choice question roughly equally often. For these 15 passages, the mean proportions of trials in which each choice was picked were similar across the three choices: M = .38 (SD = .22) for the *protagonist A*

choice, *M* = .30 (*SD* = .22) for *the protagonist B choice*, and *M* = .32 (*SD* = .28) for *the both choice*. The probability of each choice being picked did not significantly differ from chance (.33) (protagonists A choice, $t(19)= 1.15$, *n.s*; protagonists B choice, $t(19)= .53$, *n.s*; both choice, $t(19)= .32$, *n.s*).

[3] The honorific status (whether or not the referent of the subject should be respected) could be marked on the verb, but the items in this experiment did not have any honorific marking on the verbs.

Table 1. The mean proportions (SD) of trials in which participants selected the target (correct) protagonist that was indicated by the location of the gesture in the third sentence in Experiment 1 (regarding cohesive gestures).

	The gesture in the third sentence indicates the referent			
	on the left	z-score	on the right	z-score
Adults	1. 00 (0.00)	4.90***	0.99 (0.04)	4.74***
10 years	0.83 (0.35)	3.33**	0.81 (0.36)	3.19**
6 years	0.64 (0.33)	1.85*	0.69 (0.30)	2.68**
5 years	0.56 (0.31)	1.01[a]	0.52 (0.30)	0.32[b]

*** $p < .001$, ** $p < .01$, * $p < .10$ for Wilcoxon Signed-ranks test against the chance level (.05).

a p-value for 5-year-olds is 0.31.

b p-value for 5-year-olds is 0.75.

Table 2. Mean proportion (*SD*) of trials with two types of error for each age

group (out of 15 trials) in Experiment 1 (regarding cohesive gestures).

Type of error	5 years	6 years	10 years	Adults
Proportion of incorrect-protagonist choice	0.40 (0.25)	0.21 (0.16)	0.04 (0.14)	0.00 (0.01)
Proportion of both-protagonist choice	0.06 (0.21)	0.13 (0.29)	0.13 (0.34)	0.00 (0.01)

Table 3. The mean (*SD*) proportion of trials with a correct choice among the

trials with either one of the two one-protagonist choices in Experiment 1 (regarding cohesive gestures).

	M (SD)	z-score
5 years (*N*=23)	.57 (.24)	1.30
6 years (*N*=23)	.76 (.19)	3.84***
10 years (*N*=21)	.95 (.16)	4.25***
Adults (*N*=24)	1.00 (0.1)	4.81***

*** $p<.001$ for Wilcoxon Signed-ranks test against the chance level (.50)

Figure captions

Figure 1. Example stimulus used in Experiment 1 (regarding cohesive gestures). The top panel (speech): Words in boldface are accompanied by a gesture and underlines indicate periods in which a gesture(s) was held in the air. The abbreviations in the interlinear gloss are ACC (accusative), DAT (dative), NOM (nominative), PST (past tense), PROG (progressive aspect) and TOP (topic marker). "Nori" and "Yuuto" are a common Japanese boy's names. "kun" is a honorific for a young boy. The numbers in parentheses indicate where gestures occurred, and correspond to the numbers in the bottom panel. Note that Japanese does not have articles or commonly used third person pronouns; thus, it is natural to use full noun phrases for maintained referents. It allows omission of arguments as in the third sentence. The bottom panel (gesture): Gestures that accompanied the speech stimulus in the top panel.

Figure 2. Mean proportion of trials with a correct choice for each age group (out of 15 trials) in Experiment 1 (regarding cohesive gestures). The error bars indicate standard errors.

Figure 3. Example stimulus with hand-held (the Identifiable items) pictures used in Experiment 2 (regarding cohesive presentation of pictures). The vignettes were identical to Experiment 1, except that hand-held pictures replaced gestures. All symbols and abbreviations used in panels are identical to *Figure1*. The numbers in parentheses in the top panel indicate where the action with the corresponding number in the bottom panel took place.

Figure 1

1. (2)Nori-kun(3)-to (4)Yuuto-kun(5)-ga hodoukyou wo watasimasu

Nori-kun-and Yuuto-kun-NOM pedestrian bridge-ACC cross.PROG.Polite

"(2)Nori-kun (3) and (4)Yuuto-kun(5) are crossing a pedestrian bridge."

2. (6)Nori-kun(7)-to (8)Yuuto-kun-wa(9) (10)kaidan wo nobotteimasu

Nori-kun-and Yuuto-kun-TOP stairs-ACC ascend.PROG.Polite

"(6)Nori-kun (7) and (8)Yuuto-kun(9) (10)are ascending stairs."

3. Suruto totsuzen, (11)korondeshimaimashita

and suddenly tubmle.down-regrettably.Polite.PST

"and suddently, (11)tumbled down."

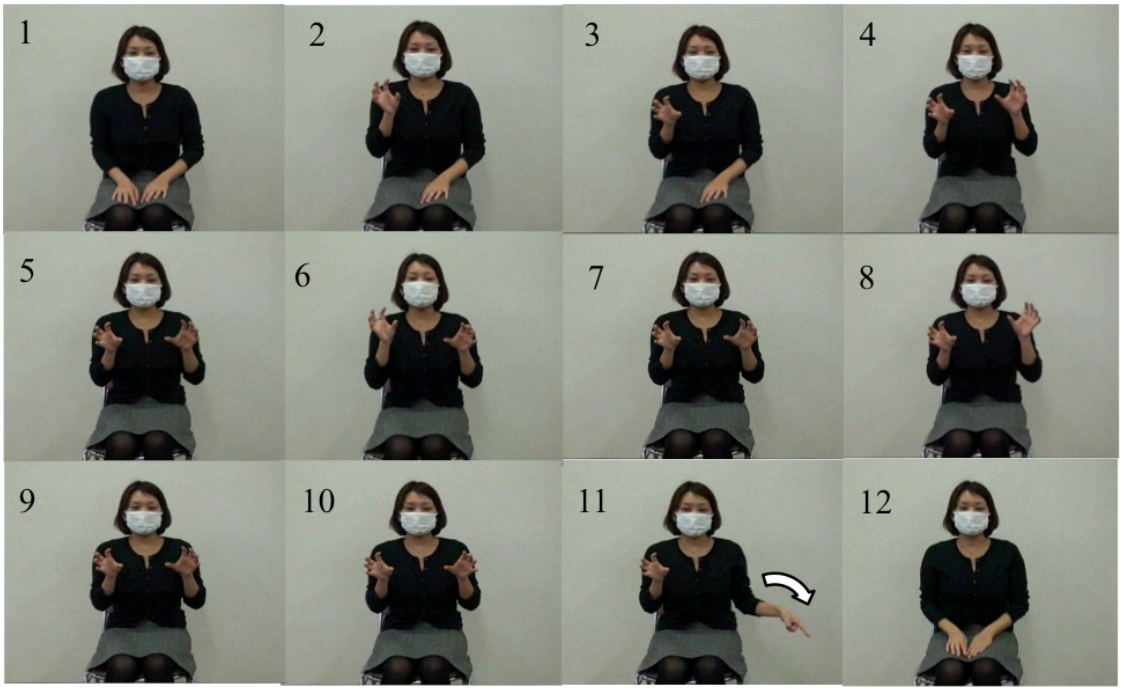

Figure 2

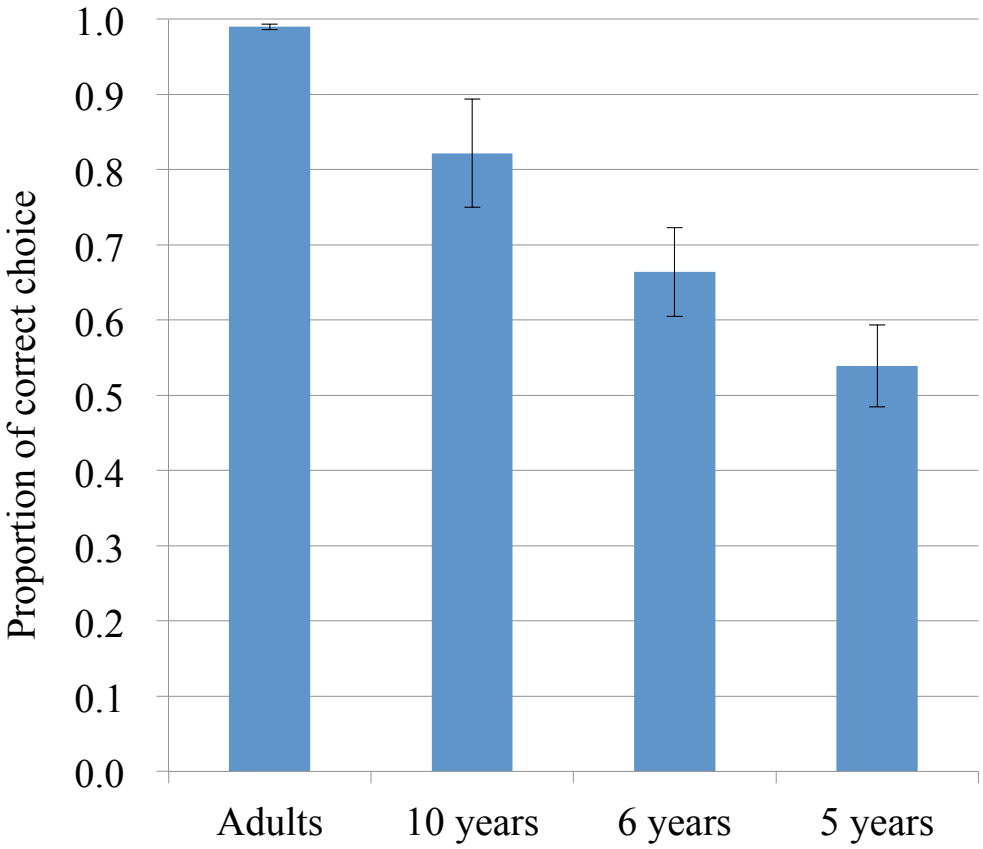
Proportion of correct choice
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Adults
10 years
6 years
5 years

Figure 3

1. (2)Nori-kun(3)-to (4)Yuuto-kun(5)-ga hodoukyou wo watasimasu

Nori-kun-and Yuuto-kun-NOM pedestrian bridge-ACC cross.PROG.Polite

"(2)Nori-kun (3) and (4)Yuuto-kun(5) are crossing a pedestrian bridge."

2. (6)Nori-kun(7)-to (8)Yuuto-kun-wa(9) (10)kaidan wo nobotteimasu

Nori-kun-and Yuuto-kun-TOP stairs-ACC ascend.PROG.Polite

"(6)Nori-kun (7) and (8)Yuuto-kun(9) (10)are ascending stairs."

3. Suruto totsuzen, (11)korondeshimaimashita

and suddenly tubmle.down-regrettably.Polite.PST

"and suddently, (11)tumbled down."